CORRELATION

Dr. D. K. Paul

Associate Professor & Course Coordinator
M. Sc. Environmental Science & Management
Department of Zoology, Patna University
Member, SEAC (for EIA), Bihar, Constituted by
MoEF & CC, Govt. of India
dkpaul.pat31@gmail.com

Introduction

- When two series of measurements are made at the same time or on the same organism, it is frequently desirable to know whether a relationship exists between variables or not.
- A relationship between two such sets of measurements such that value of one measurement is affected by the value of other is described as correlation and measurement of such relationship is called correlation analysis.
- Relationship between two variable's , the correlation analysis is called bivariate analysis and analysis of more than two variables is called multivariate analysis.

DEFINITION

1. The relationship between the two variables are such that a change in one variable results in a change in the other variable is known as correlation.
2. According to Craxton and Cowden " When the relationship is of quantitative nature , the appropriate statistical tool for discovering and measuring the relationship and expressing it in a brief formula is known as "correlation".
3. Correlation analysis deals with the association between two or more variables – Simpson and Kafka
4. Correlation analysis attempts to determine the degree of relationship between variables- Ya Lun Chou

Significance of the Study of Correlation

The study of correlation is of immense use in practical life because of the following reasons:

1. Most of the variables have some kind of relation ex. Hb and number of RBC.
2. Once we know that the variable are closely related, we can estimate the value of one variable given the value of another (By regression analysis).
3. Correlation analysis contribute to the understanding of economic behaviour.
4. Correlation analysis is likely to be more valuable and near to reality.

Correlation and Causation

Correlation analysis helps us in degerming the degree of relationship between two or more variables. It does not tell us anything about cause and effect relationship. Even a high degree of correlation does not necessarily mean that a relationship of cause and effect exits between the variables. The explanation of a significant degree of correlation may be any one or combination of following reasons:

1. The correlation may be due to pure chance especially in a small sample.
2. Both the correlated variables may be influenced by one or more other variables.
3. Both the variables may be mutually influencing each other, so that non can be designated as the cause and the other the effect

Types of Correlation

There are three kinds of correlation:

A. Positive Correlation

B. Negative Correlation

C. Zero Correlation

A. Positive or direct Correlation: When one variable increases (decreases) and the other also increases (decreases) , they are said to be positively correlated. It is of two types:

(i) Positive perfect correlation: In this, the two variables denoted by X and Y are directly proportional and fully correlated with each other . The value of correlation coefficient (r) in case of perfect positive correlation is 1. However, in biological sciences ,it is rare.

Types of Correlation

ii. Moderately positive correlation: The two variables denoted by X and Y are partially correlated with each other ex. Age of husband and age of wife . The value of correlation coefficient in case of moderate positive correlation (r) lies between 0 and 1. It gives scatter diagram.

B. Negative or Inverse Correlation : When one variable increases (decreases) and the other decreases (increases) , they are said to be negatively correlated. It is of two types:

(i) Negative perfect correlation: In this, the two variables denoted by X and Y are inversely proportional and negatively correlated with each other . The value of correlation coefficient (r) in case of perfect positive correlation is -1. However, it is rare in Life science.

Types of Correlation

ii. Moderately negative correlation: The two variables denoted by X and Y are partially correlated with each other inversely ex. Age and vitality in adults. The value of correlation coefficient in case of moderate positive correlation (r) lies between 0 and -1. It also gives scatter diagram.

C. Zero Correlation: If the variation in one variable has no relation with the variation in the other, the two variables has no correlation or zero correlation ex. body weight and I.Q.

Curvilinear Correlation: In addition to the above linear correlation , there is also curvilinear correlation. The correlation of scores of two variables based on some quality shown by a curve on graph is called curvilinear correlation.

Ex. The correlation between practice and degree of learning is positive curvilinear correlation

Time and degree of retention (memory) shows negative curvilinear correlation

Properties of Coefficient of Correlation

1. Correlation coefficient (r) lies between -1 and +1 i.e. $-1 \leq r \geq 1$.
2. If $r = +1$, the correlation is perfect positive.
3. If $0 < r < 1$, the correlation is moderately positive.
4. If $r = -1$, the correlation is perfect negative.
5. If r lies between 0 and -1, then the correlation is moderately negative.
6. If $r = 0$, , there is no correlation between two variables.

METHODS OF STUDYING CORRELATION

There are various methods of studying correlation between two variables .Among those , following methods are popular:

A. Scatter diagram

B. Karl Pearson's Coefficient of Correlation

C. Rank Correlation

A. Scatter diagram

Scatter diagram is a graphical method of showing the correlation between the two variables . Let X_1, Y_1 (i = 1,2,3,… n) be a bivariate. The values of the variable X are plotted along the OX-axis corresponding to every ordered pair (X_1, Y_1). We have a point in the co-ordinate plane. The diagram obtained by plotting these n points is called the scatter or dot diagram.

METHODS OF STUDYING CORRELATION

Merits and demerits of scatter diagram:

- Scatter diagram method is simple to find out the nature of correlation between the two variable.
- It is the first step to find out the relationship between two variable.
- This method gives the idea about the direction of correlation.
- It also gives an idea about the nature of correlation , whether it is positive or negative .
- But, we cannot get the exact degree of correlation, between the two variables, as it is only possible by the coefficient of correlation.
- This method is not influenced by the extreme items.

METHODS OF STUDYING CORRELATION

(B) Karl Pearson’s Coefficient of Correlation

- When the relationship between two sets of variables is described by a straight line, then the correlation between variables may be expressed by the ‘product moment’ coefficient of correlation ($r_{x.y}$ or $r_{y.x}$) .
- The degree or the extent to which the variables of a bivariate distribution are related with each other is called the coefficient of correlation ,designated by the letter “r”.
- Karl Pearson (1867-1936) suggested a mathematical method for measuring the closeness of relationship between two variables. Pearson’s measure known as Pearson’s product moment correlation coefficient between two variables X an Y , usually denoted by r (X,Y) or $r_{x.y}$ or simply ‘r’.

METHODS OF STUDYING CORRELATION

- According to J P Guilford “ Coefficient of correlation is a single number that ells us to what extent two things are related and to what extent variations in one go with variations in other”.
- Karl Pearson’s Coefficient of correlation is given by the expression as follows:

$$r_{x.y} = \frac{\sum x.y}{\sqrt{\sum x^2 . \quad \sum y^2}}$$

Where r = correlation coefficient, x = deviations of X variables, y= deviations of Y variables

- In language , we can say that ’r’ can be calculated by dividing the sum of products of deviations from the square root of the products of the sums of squares of deviations of the two variables.

METHODS OF STUDYING CORRELATION

Example : The length and weight of 7 groups of fishes of a species is given below. Find the correlation coefficient of the two variables.

Length (in cm)	11.7	13.9	15.5	17.8	18.5	19.2	21.0
Weight (in gm)	7.10	12.42	15.35	23.20	28.45	32.25	39.84

S.N	Length X	Weigth Y	X- $\bar{X}$ x	Y- $\bar{Y}$ y	X^2	y^2	x.y
1	11.7	7.10	-5.1	-15.55	26.01	241.8	79.3
2	13.9	12.42	-2.9	-10.23	8.41	104.6	29.66
3	15.5	15.35	-1.3	-7.3	1.69	53.2	9.49
4	17.8	23.20	1.0	0.55	1.0	0.30	0.55
5	18.5	28.45	1.7	5.8	2.89	33.64	9.86
6	19.2	32.25	2.4	9.6	5.76	92.16	23.04
7	21.0	39.84	4.2	17.19	17.64	295.49	72.2
N=7	$\sum X$ =117.6	$\sum Y$ =158.6			$\sum X^2$=63.4	$\sum Y^2$=821.19	$\sum X$.Y=224.1

METHODS OF STUDYING CORRELATION

$\bar{X} = \frac{\sum X}{N} = \frac{117.6}{7} = 16.8$ $\bar{Y} = \frac{\sum Y}{N} = \frac{158.6}{7} = 22.65$

$$r_{x.y} = \frac{\sum x.y}{\sqrt{\sum x^2 . \quad \sum y^2}} = \frac{224.10}{\sqrt{63.4 \;\; X\; 821.19}}$$

$$= \frac{224.10}{\sqrt{52063.45}}$$

$$= \frac{224.10}{228.17} = 0.98$$

$r_{x.y} = 0.98$

METHODS OF STUDYING CORRELATION

- Conclusion: There is a strong positive correlation between the length and weight of body of studied species of fish. Calculated value comes to 0.98. So, r is very nearest to 1 . Therefore, both variables are highly correlated.
- Calculation of ‘r’ using raw scores (Direct method) :

Correlation coefficient ‘r’ can be find out with the help of following formula using raw data without finding deviation.

$$r= \frac{\left[\sum X \cdot Y - \frac{\sum X \cdot \sum Y}{N}\right]}{\sqrt{\left\{\Sigma X^2 - \frac{(\sum X)^2}{N}\right\} X \left\{\Sigma Y^2 - \frac{(\sum Y)^2}{N}\right\}}}$$

This is a simple and direct method like the one used to find the sum of squares in standard deviation. Mean is not required.

METHODS OF STUDYING CORRELATION

- Example: On entry to a school, a new intelligence test was given to a small group of children. The results obtained in that test and in a subsequent examination are tabulated as follow. Find out the correlation coefficient.

Student	1	2	3	4	5	6	7	8
Intelligence Test score	6	4	6	8	8	10	8	6
Marks in exam	4	4	7	10	4	7	7	1

METHODS OF STUDYING CORRELATION

Child number	Intelligent Test score		Marks in exam		
	X	X^2	Y	Y^2	X.Y
1	6	36	4	16	24
2	4	16	4	16	16
3	6	36	7	49	42
4	8	64	10	100	80
5	8	64	4	16	32
6	10	100	7	49	70
7	8	64	7	49	56
8	6	36	1	1	06
N=8	$\sum X$=56	$\sum X^2$=416	$\sum Y$=44	$\sum Y^2$=56	$\sum X.Y$=326

METHODS OF STUDYING CORRELATION

$$r=\frac{\left[\sum X\cdot Y-\frac{\sum X\cdot\sum Y}{N}\right]}{\sqrt{\left\{\Sigma X^{2}-\frac{(\sum X)^{2}}{N}\right\}X\left\{\Sigma Y^{2}-\frac{(\sum Y)^{2}}{N}\right\}}}$$

$$=r=\frac{\left[326-\frac{56X44}{8}\right]}{\sqrt{\left\{416-\frac{56^{2}}{8}\right\}X\left\{296-\frac{44^{2}}{8}\right\}}}=\frac{[326-308]}{\sqrt{\{416-392\}X\{296-242\}}}$$

$$=\frac{18}{\sqrt{24X54}}=\frac{18}{36}=0.5$$

r=0.5 Ans.

METHODS OF STUDYING CORRELATION

- To interpret the observed correlation coefficient (0.5) , the test of significance is applied i.e. finding the standard error (SE) of correlation which is a measure of sampling variation.
- Correlation coefficient gives a measure of the degree of relationship between the two variables of a sample. But it does not indicate that whether 'r' is larger enough to suggest the existence of a correlation between paired values in the population or whether the obtained 'r' can be attributed to coincidence in order to test the significance of correlation coefficient 'r' from zero. One may apply the 't' test with n-2 degree of freedom. The formula of 't' test is as follow

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-(r)^2}} = \frac{0.5\sqrt{6}}{\sqrt{1-(0.5)^2}} = \quad 1.414$$

METHODS OF STUDYING CORRELATION

- For n=2 i.e. 6 degree of freedom at 0.5 level of significance , the highest value of ‘t’ obtainable by chance is 2.447. The calculated value of ‘t’ is 1.414. It is less than table value i.e. 2.447. Hence it is not significant at 0.05 level of significance.
- The hypothesis of no correlation is accepted and correlation between intelligence score and marks in examination is not established.
- Decision on the Karl Pearson’s coefficient table: Significance of this sample may also be found by direct reference to correlation coefficient table. For 6 degree of freedom , highest value of ‘r’ is 0.707 at 0.05 level of significance. Our calculated value is 0.5, hence it is insignificant at 5% level ($p > 0.05$).
- Conclusion: The new intelligence test is not useful enough to predict the performance in examination though apparently it may appear to be so.
- In a large sample even low degree of correlation ‘r’ may be highly significant while in a small sample, high degree of r may be insignificant.

Merits and Demerits of the Pearson's Coefficient

- Amongst the mathematical methods used for measuring the degree of relationship, Karl Pearson's method is most popular. The correlation coefficient summarizes in one figure not only the degree of correlation but also be the direction i.e. whether correlation is positive or negative.
- However there some limitations in this method. The chief limitations of the method are:

I. The correlation coefficient always assumes linear relationship regardless of the fact whether that assumption is correct or not.

II. Great care must be exercised in interpreting the value of this coefficient as very often the coefficient is misinterpreted.

III. The value of the coefficient is widely affected by the extreme values.

IV. As compared with other methods, this method take more time to compute the value of correlation coefficient.

Rank Correlation

- The Karl Pearson's method is based on the assumption that the population being study is normally distributed. When it is known that the population is not normal or when the shape of the distribution is not known, there is need for a measure of correlation that involves no assumption about the parameter of the population.
- If the variables are not capable of quantitative measurement but arranged in derail order for example color, fecundity, stress tolerance etc. and which can not be measures quantitively but can be arranged serially, Rank's correlation method is applied.
- Charles Edward Spearman (1904) developed a formula which helps in obtaining the correlation coefficient between the ranks of n individuals in two characteristic under study.

Rank Correlation

- Spearman's rank correlation coefficient between the usually denoted by ρ (rho) is given by the formula

- R or $\rho = 1 - \frac{6\Sigma D^2}{n(\mathrm{n}^2-1)}$

Where,

R or ρ = rank difference of X and Y variables

D = difference of ranks between the [pairs of same individuals in the two characteristics

n = number of pairs

$\sum \mathrm{D}^2 =$ Summation of square of difference of two variables rank 1 and 2 (R_1& R_2)

Rank Correlation

- Example (Ranks are given):

Two ladies Neelu and Reena were asked .to rank 7 different types of lipsticks . The ranks given by them are as follow

Lipsticks	A	B	C	D	E	f	G	
Neelu		2	1	4	3	5	7	6
Reena		1	3	2	4	5	6	7

Calculate Spearman's rank correlation coefficient.

Soln.

S.N.	X R_1	Y R_2	R_1-R_2 i.e. D	D^2
A	2	1	1	4
B	1	3	-2	4
C	4	2	2	4
D	3	4	-1	1
E	5	5	0	0
F	7	6	1	1
G	6	7	-1	1
				$\sum D^2 = 12$

Rank Correlation

$$R = 1 - \frac{6\Sigma D^2}{n(\mathrm{n}^2-1)}$$

$$= 1 - \frac{6\ X\ 12}{7(7^2-1)}$$

$$= 1 - \frac{72}{7\ X\ 48}$$

$$= 1 - \frac{3}{14} = \frac{11}{14} = 0.786$$

Calculated R = 0.786. Table value of R by Spearman’s Rank difference correlation at N(7) is 0.745 at 0.05 level of significance.

Hence it is significant.

Rank Correlation

- When Ranks are not given

When we are given the actual data and not the ranks, it will be necessary to assign the ranks. Ranks are assigned by taking either highest value as 1 or the lowest value as 1. But whether we start the lowest value or the highest value we must follow the same method in case of both the variable.

Example : Calculate Spearman's coefficient of correlation between marks assigned to ten students by judges X and Y in a certain competition test as shown below:

S.N	1	2	3	4	5	6	7	8	9	10
Marks by judge X	52	53	42	60	45	41	37	38	25	27
Marks by judge Y	65	68	43	38	77	48	30	32	25	50

Rank Correlation

S.N	Marks by judge X	Rank X Rx	Marks by judge Y	Rank Y Ry	Rx-Ry=D	D^2
1	52	8	65	8	0	0
2	53	9	68	9	0	0
3	42	6	43	5	1	1
4	60	10	38	4	6	36
5	45	7	77	10	-3	9
6	41	5	48	6	-1	1
7	37	3	30	2	1	1
8	38	4	32	3	1	1
9	25	1	25	1	0	0
10	27	2	50	7	-5	25 $\sum \mathrm{D}^2 = 74$

Rank Correlation

R $= 1 - \frac{6\Sigma D^2}{n(\mathrm{n}^2-1)}$ (putting the values)

$= 1 - \frac{6\ X\ 74}{10(10^2-1)}$

$=1 - \frac{444}{10(100-1)}$

$=1 - \frac{444}{10(99)}$

$=1 - \frac{444}{990}$

$=\frac{546}{990}$

= 0.5515

R=0.5515

Rank Correlation

- Decision: Table value of R for N=10 is 0.6364 (at 0.05 level) and 0.7818 (at 0.01 level). Calculated value 0.5515 of R is less than the table value. Hence it is not significant.
- Inference :Marks given by the judges are so much differ.

Suggested Readings

- Zar JH(2009). Biostatistical Analysis. Dorling Kindersley Pvt. Ltd., India.
- Selvin S (2007). Biostatistics how it Works. Pearson Education.
- Mahajan BK (1999) .Methods in Biostatistics. Jaypee Brothers, New Delhi.
- Sunder Rao PSS and Richard J (2001). An Introduction To Biostatistics. Prentice Hall of India Private Limited. New Delhi.
- Prasad S (2007). Elements of Biostatistics. Rastogi Publications, Meerut.

THANKS